2016

Statistical properties of atmospheric greenhouse gas measurements: looking down from space and looking up from the ground

Bohai Zhang
University of Wollongong

Noel Cressie
University of Wollongong

Debra Wunch
University of Toronto

Statistical properties of atmospheric greenhouse gas measurements: looking down from space and looking up from the ground

Abstract

Remote sensing platforms can collect measurements on a global scale within a few days, which provides an unprecedented opportunity to characterize and understand the spatio-temporal variability of environmental variables. Because of the additional challenges of making precise and accurate measurements from space, it is essential to validate satellite remote sensing datasets with highly precise and accurate ground-based measurements. The focus of this article is on two sets of measurements: Atmospheric column-averaged carbon dioxide (CO 2) collected by the Orbiting Carbon Observatory-2 (OCO-2) mission in its target mode of operation; and ground-based data used for validation from the Total Carbon Column Observing Network (TCCON). The current statistical modeling of the relationship between the OCO-2 data and the TCCON data assumes a linear regression and different measurement errors that reside in both the TCCON data and the OCO-2 data. To obtain consistent estimates of the regression coefficients, it is critical to determine the error variance of each datum in the regression. In this article, a rigorous statistical procedure is presented for obtaining the error variances through modeling the spatial and/or temporal dependence structure in the OCO-2 and TCCON datasets. Numerical results for analyzing a pair of datasets at the Lamont TCCON station and OCO-2 orbit number 3590 illustrate our procedure.

NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

08-16

Statistical Properties of Atmospheric Greenhouse Gas Measurements: Looking Down from Space and Looking Up from the Ground

Bohai Zhang, Noel Cressie, and Debra Wunch

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522, Australia Phone +61 2 4221 5435, Fax +61 2 4221 4998.
Email: karink@uow.edu.au

Statistical properties of atmospheric greenhouse gas measurements: looking down from space and looking up from the ground

Bohai Zhang [a], Noel Cressie [a], Debra Wunch [b]

[a] *National Institute for Applied Statistics Research Australia, University of Wollongong, NSW, 2522, Australia*

[b] *Physics and School of the Environment, University of Toronto, Toronto, ON, M5S 3E8, Canada*

Abstract

Remote sensing platforms can collect measurements on a global scale within a few days, which provides an unprecedented opportunity to characterize and understand the spatio-temporal variability of environmental variables. Because of the additional challenges of making precise and accurate measurements from space, it is essential to validate satellite remote sensing datasets with highly precise and accurate ground-based measurements. The focus of this article is on two sets of measurements: Atmospheric column-averaged carbon dioxide (CO_2) collected by the Orbiting Carbon Observatory-2 (OCO-2) mission in its target mode of operation; and ground-based data used for validation from the Total Carbon Column Observing Network (TCCON). The current statistical modeling of the relationship between the OCO-2 data and the TCCON data assumes a linear regression and different measurement errors that reside in both the TCCON data and the OCO-2 data. To obtain consistent estimates of the regression coefficients, it is critical to determine the error variance of each datum in the regression. In this article, a rigorous statistical procedure is presented for obtaining the error variances through modeling the spatial and/or temporal dependence structure in the OCO-2 and TCCON datasets. Numerical results for analyzing a pair of datasets at the Lamont TCCON station and OCO-2 orbit number 3590 illustrate our procedure.

Key words: atmospheric carbon dioxide, errors in variables, OCO-2, spatial statistics, TCCON, time series

1 Introduction

Satellite remote sensing measurements of Earth's surface and atmosphere provide global coverage within a matter of days. This helps scientists understand the spatio-temporal distribution of environmental processes. Examples of remote sensing datasets of this type include atmospheric trace gases (e.g., carbon dioxide, methane, ozone), sea surface temperature, sea-ice extent, solar-induced fluorescence from plants, aerosols, and so forth. These remote sensing measurements from space require validation from well characterized ground-based measurements to ensure their accuracy and precision throughout the satellite's mission. This is often achieved through fitting a regression relationship between coincident ground-based and satellite-based measurements, where uncertainties are present in both the dependent (Y) and independent (X) variables. Consequently, correctly determining the error variance of each of the X- and Y-values in the regression is of critical importance.

In this article, we determine the error variances of these values by exploiting spatial dependence for the satellite-based measurements and temporal dependence for the ground-based measurements. Finding these variances is key to estimating accurately the regression parameters. The problem is generally relevant to many topics in chemistry, physics, and the biogeosciences; in this article, we shall focus on the Orbiting Carbon Observatory-2 (OCO-2) validation program [1].

The OCO-2 mission aims to provide the atmospheric measurements required to understand better the carbon cycle, which is the cycling of carbon (most often in the form of CO_2) between the oceans, land, terrestrial biosphere, and atmosphere. The main sinks of CO_2 are the oceans, which dissolve CO_2 into seawater to form carbonic acid, and the terrestrial biosphere, in which plants, through photosynthesis, convert CO_2 into the sugars necessary to grow [2]. There are many sources of atmospheric CO_2, primarily fossil-fuel burning (e.g., coal, petroleum, natural gas), which oxidizes carbon-containing fuels to produce CO_2; and land use, which both alters the surface albedo and on average reduces the land CO_2 sink [2]. Other industrial processes are significant sources of CO_2, such as cement production, where limestone ($CaCO_3$) is chemically converted into calcium oxide (CaO), producing CO_2 as a byproduct [2,3].

Due to these human activities, CO_2 concentrations in Earth's atmosphere have been increasing: The atmospheric CO_2 concentration of Earth has increased from about 280 parts per million (ppm) since the beginning of the industrial revolution in the 1700s to about 400 ppm today. The percentage of each year's CO_2 emissions that remain in the atmosphere has also been increasing in the past 50 years. According to [4], there is evidence that from 1959 to 2008,

the fraction of CO_2 emissions that remain in atmosphere each year is likely to have increased from 40% to 45%. The increasing levels of atmospheric CO_2 and other greenhouse gases are the primary cause of increases in Earth's surface temperature.

The goals of the OCO-2 mission are to measure atmospheric carbon dioxide (CO_2) with high enough precision and accuracy to distinguish between the sources and sinks of CO_2 on regional scales, and to quantify the seasonal, latitudinal, and interannual variability of CO_2 [5]. To achieve this goal, measurement precision and accuracy must be better than one part per million of CO_2 (i.e., 1 ppm or 0.25%) [6]. This is a difficult task and, thus, the method to ensure that the OCO-2 data are sufficiently accurate is critical. The standard CO_2 gas scale is set by the World Meteorological Organization (WMO), and to tie the OCO-2 data to that standard scale requires a so-called transfer standard between the WMO-calibrated instruments and the remote sensing OCO-2 measurements. The Total Carbon Column Observing Network (TCCON, [7]) acts as this transfer standard, since TCCON is tied to the WMO scale through comparisons with WMO-traceable aircraft and balloon-borne measurements [8].

To compare OCO-2 data with TCCON data, a special observation mode was designed for the OCO-2 satellite, called "target mode." In this mode, the OCO-2 spacecraft turns to "stare" at a ground location (typically a TCCON station) as it passes overhead, recording thousands of measurements in a small geographic area ($\sim 0.2 \times 0.2$ degrees) over just a few minutes (~ 5 minutes). Under these conditions, changes in atmospheric CO_2 abundances are negligibly small, and the OCO-2 data obtained from looking down from space are directly coincident and comparable with the TCCON data obtained from looking up from the ground [9]. After some preprocessing of the OCO-2 data, a linear regression between OCO-2 and TCCON is fitted, and the deviation of the OCO-2 CO_2 product from the WMO scale is quantified [10]. This deviation is removed from the OCO-2 data, using the fitted regression line, before its use in scientific studies. Thus, fitting this regression line correctly is imperative, which we show in this article depends on the error variance of each OCO-2 and TCCON value used in the regression.

The result we obtain for the variance of the OCO-2 value in the regression can be used in contexts that go beyond this calibration study. For example, flux inversions usually work with spatially aggregated mole-fraction data (e.g., from OCO-2 retrievals), and our research demonstrates how the spatial covariance of the mole-fraction field determines the all-important aggregated variances. Another benefit of our research is for small area analysis [10–12], where we are able to account for dependence in the constituent random variables. This statistical dependence results in a modification (often reduction) of the number of observations, which we call the "effective" number of observations. The

interpretation of these, in terms of reduced information content in the small areas, is powerful and intuitive.

The rest of the paper is organized as follows. In Section 2, we provide the details of the errors-in-variables model that is currently used by the OCO-2 validation team. In Section 3, we discuss the selection of weights in their regression of OCO-2 on TCCON, and we provide sufficient conditions for obtaining unbiased estimating equations of regression parameters. In this section, we also illustrate through simulation the benefits of using an unbiased estimating equation for the regression slope b. We elaborate the statistical-analysis procedures for individual TCCON and OCO-2 datasets in Section 4, focusing on modeling the temporal and spatial data-dependence structures of the TCCON and OCO-2 datasets, respectively. In Section 5, we provide the formulas for computing the variances of the values fitted in the regression, and we give the Lamont TCCON site and OCO-2 orbit number 3590 as an example. Concluding remarks are given in Section 6, and the paper finishes with a technical appendix.

2 The errors-in-variables model used for OCO-2 calibration

Version 7 of the OCO-2 retrieval data product is publicly available and can be found at [13]. The regression procedure used to obtain version 7 is described in [10], as follows. Let (X_i, Y_i) be a pair of TCCON and OCO-2 target-mode observations, where i indexes a combination of station and OCO-2 orbit number. Suppose there are $i = 1, \ldots, N$ such combinations. The errors-in-variables model in [14,15] was used to model the linear relationship between these pairs, with TCCON as the independent variable (X) and OCO-2 as the dependent variable (Y). An iterative algorithm in [15] was used to estimate the regression coefficients.

In this article, we show that the weighted-least-squares estimators of the regression coefficients in [14,15] are (asymptotically) unbiased and (statistically) efficient when the regression weights are properly specified as being inversely proportional to $\mathrm{var}(X_i)$ and $\mathrm{var}(Y_i)$, respectively, for $i = 1, \ldots, N$. We show that misspecified weights result in biased estimating equations of the regression parameter, and hence the resulting regression-parameter estimates and regression line can be biased (Section 3). TCCON datasets are weakly correlated in time and OCO-2 datasets are highly correlated in space, which must be accounted for when estimating the variances of X_i and Y_i, respectively (Section 4).

Generally, suppose that $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)\}$ are N pairs of ground-monitoring-station data (X) paired with satellite remote sensing data Y. Since

the data collected by both the satellite and the ground-monitoring stations have measurement errors associated with them, an errors-in-variables model is appropriate for modeling their relationship. Assume that $E(X_i) = x_i$ and $E(Y_i) = y_i$, where x_i and y_i are (unknown) true values of X_i and Y_i, respectively. Because both datasets attempt to measure the same variable (e.g., in our application, column-averaged CO_2), it is expected that there is a strong relationship between them. Fitting a linear relationship provides a straightforward way for correcting the bias in the satellite data using the more accurate data from the ground-monitoring stations.

In [10], the current errors-in-variables model for producing version 7 of the OCO-2 retrieval data product is given as follows. For $i = 1, \ldots, N$,

$$\begin{aligned} X_i &= x_i + \epsilon_{x,i}, \\ Y_i &= y_i + \epsilon_{y,i}, \\ y_i &= a + bx_i, \end{aligned} \tag{1}$$

where the measurement-error terms, $\epsilon_{x,i}$ and $\epsilon_{y,i}$, are assumed to have mean zero and variances $\tilde{\sigma}^2_{x,i}$ and $\tilde{\sigma}^2_{y,i}$, respectively. It is also reasonable to assume that they are mutually independent for all $i = 1, \ldots, N$.

It is important to clarify that X_i and Y_i are aggregated data calculated from a set of individual TCCON observations and a set of individual OCO-2 target-mode data, respectively. Let $n_{x,i}$ and $n_{y,i}$ denote sample sizes of individual TCCON and OCO-2 observations for obtaining X_i and Y_i, respectively. In this article, we derive

$$\tilde{\sigma}^2_{x,i} \equiv \mathrm{var}(X_i) = \sigma^2_{x,i}/\tilde{n}_{x,i}, \qquad \tilde{\sigma}^2_{y,i} \equiv \mathrm{var}(Y_i) = \sigma^2_{y,i}/\tilde{n}_{y,i}, \tag{2}$$

where $\sigma^2_{x,i}$ and $\sigma^2_{y,i}$ are variances of a single TCCON observation and a single OCO-2 observation, respectively; and $\tilde{n}_{x,i}$ and $\tilde{n}_{y,i}$ are the effective sample sizes, respectively (which are different from $n_{x,i}$ and $n_{y,i}$). Strong positive dependence between individual data results in effective sample sizes much smaller than actual sample sizes (Section 5).

A least-sum-of-weighted-squares criterion was proposed in [16]: This leads to estimating a and b by minimizing

$$S(a, b) = \sum_{i=1}^{N} \frac{w_{x,i} w_{y,i}}{b^2 w_{y,i} + w_{x,i}} (Y_i - a - bX_i)^2, \tag{3}$$

with respect to a and b. The resulting estimates, $\hat{a}_{lws}$ and $\hat{b}_{lws}$, are functions only of the data $\{(X_i, Y_i)\}_{i=1}^N$, for given regression weights $\{w_{x,i}\}_{i=1}^N$ and

$\{w_{y,i}\}_{i=1}^N$.

Partially differentiating (3) with respect to a and b, one can obtain (see [16]),

$$b = \frac{\sum_{i=1}^{N} Z_i^2 (Y_i - \bar{Y}_w)\left(\frac{X_i - \bar{X}_w}{w_{y,i}} + \frac{b(Y_i - \bar{Y}_w)}{w_{x,i}}\right)}{\sum_{i=1}^{N} Z_i^2 (X_i - \bar{X}_w)\left(\frac{X_i - \bar{X}_w}{w_{y,i}} + \frac{b(Y_i - \bar{Y}_w)}{w_{x,i}}\right)}, \tag{4}$$

where $Z_i \equiv (w_{x,i}w_{y,i})/(b^2 w_{y,i} + w_{x,i})$ depends on b, $\bar{X}_w = \sum_i Z_i X_i / \sum_i Z_i$, and $\bar{Y}_w = \sum_i Z_i Y_i / \sum_i Z_i$. Based on equation (4), [15] proposed an algorithm that solves for b iteratively. If it converges, the resulting estimate is $\hat{b}_{lws}$. Then the corresponding estimate of a is $\hat{a}_{lws} = \bar{Y}_w - \hat{b}_{lws}\bar{X}_w$, where $\bar{X}_w$ and $\bar{Y}_w$ are evaluated at $b = \hat{b}_{lws}$.

3 Unbiased estimation of regression parameters

In this section, we find sufficient conditions under which the least-sum-of-weighted-squares estimators, $\hat{a}_{lws}$ and $\hat{b}_{lws}$, are (asymptotically) unbiased. The estimating equations for regression parameters a and b are $\frac{\partial S(a,b)}{\partial a} = 0$ and $\frac{\partial S(a,b)}{\partial b} = 0$, respectively; the estimating equations are unbiased if $E(\frac{\partial S(a,b)}{\partial a}) = 0$ and $E(\frac{\partial S(a,b)}{\partial b}) = 0$, for all $a, b \in \mathbb{R}$. Under regularity conditions, unbiased estimating equations result in consistent (asymptotically unbiased) estimators [17,18]. Therefore, unbiasedness of estimating equations of a and b is a desirable property.

Recall that $\{w_{x,i}\}$ and $\{w_{y,i}\}$ are pre-specified constants; then

$$\frac{\partial S(a,b)}{\partial a} = \sum_{i=1}^{N} \frac{-2w_{x,i}w_{y,i}}{b^2 w_{y,i} + w_{x,i}}(Y_i - a - bX_i),$$

and hence $E\left(\frac{\partial S(a,b)}{\partial a}\right) = 0$. Further,

$$\begin{aligned}\frac{\partial S(a,b)}{\partial b} = \sum_{i=1}^{N} \frac{-2w_{x,i}w_{y,i}}{(b^2 w_{y,i} + w_{x,i})^2}\{&b(Y_i - a)^2 w_{y,i} + X_i(Y_i - a)w_{x,i} \\ &-b^2 X_i(Y_i - a)w_{y,i} - bX_i^2 w_{x,i}\},\end{aligned}$$

and hence

$$E\left(\frac{\partial S(a,b)}{\partial b}\right) = \sum_{i=1}^{N} \frac{-2bw_{x,i}w_{y,i}}{(b^2w_{y,i} + w_{x,i})^2}(\tilde{\sigma}^2_{y,i}w_{y,i} - \tilde{\sigma}^2_{x,i}w_{x,i}), \tag{5}$$

since $E(X_i^2) = \tilde{\sigma}^2_{x,i} + x_i^2$, $E((Y_i - a)^2) = \tilde{\sigma}^2_{y,i} + b^2x_i^2$, and $E(X_i(Y_i - a)) = bx_i^2$. Recall that $\tilde{\sigma}^2_{x,i} = \text{var}(X_i)$ and $\tilde{\sigma}^2_{y,i} = \text{var}(Y_i)$.

From (5), if

$$\frac{w_{x,i}}{w_{y,i}} = \frac{\tilde{\sigma}^2_{y,i}}{\tilde{\sigma}^2_{x,i}}, \tag{6}$$

then $E\left(\frac{\partial S(a,b)}{\partial b}\right) = 0$, for all $a, b \in \mathbb{R}$. The solution to (5) is $\hat{b}_{lws}$, which is consistent, provided (6) holds. That is, provided the ratio of the weights associated with X_i and Y_i are equal to the reciprocal of the ratio of their corresponding true variances, $\hat{b}_{lws}$ is consistent. In the following subsection, we use a simulation example to show that when the regression weights are misspecified, the least-sum-of-weighted-squares estimator of b is biased with a large mean squared error (MSE).

3.1 Effects of regression weights on estimating the slope parameter b

To illustrate the effect of the regression weights on estimating b, consider the following artificial example based on simulation, where the units of X and Y are arbitrary and not related to our application to CO_2 mole fraction. We first generated the true covariate values $\{x_i\}_{i=1}^N$ from a Gaussian distribution, $N(10, 2^2)$, and we set the true response values $\{y_i\}_{i=1}^N$ to be given by: $y_i = 0.8x_i$. That is, the true value of a is 0 and the true value of b is 0.8. Then X_i was randomly generated from a Gaussian distribution with mean x_i and variance $\tilde{\sigma}^2_{x,i} = 0.5$, while Y_i was randomly generated from a Gaussian distribution with mean y_i and variance $\tilde{\sigma}^2_{y,i} = 1.5$. This was repeated independently for $i = 1, \ldots, N$. Thus, in the simulation, the true ratio of $\text{var}(Y_i)$ to $\text{var}(X_i)$ is $\tilde{\sigma}^2_{y,i}/\tilde{\sigma}^2_{x,i} = 3$, for all $i = 1, \ldots, N$. We estimated b under four different scenarios: 1) $w_{x,i} = 1/0.5, w_{y,i} = 1/1.5$, corresponding to the ideal case that specifies the weights of X_i and Y_i to be the reciprocals of their respective true variances ("True"); 2) $w_{x,i} = 1, w_{y,i} = 1/3$, corresponding to misspecification of the weights but correct specification of their ratio ("Equal"); 3) $w_{x,i} = 1/2, w_{y,i} = 2$, corresponding to a misspecification of the ratio, where $w_{x,i}/w_{y,i} = 1/4$ is smaller than the actual ratio, $\tilde{\sigma}^2_{y,i}/\tilde{\sigma}^2_{x,i} = 3$ ("Smaller"); and 4) $w_{x,i} = 1, w_{y,i} = 1/10$, corresponding to a misspecification of the ratio, where $w_{x,i}/w_{y,i} = 10$ is bigger than the actual ratio, $\tilde{\sigma}^2_{y,i}/\tilde{\sigma}^2_{x,i} = 3$ ("Bigger"). The slope parameter b was estimated by minimizing the sum-of-weighted-squares objective function in (3).

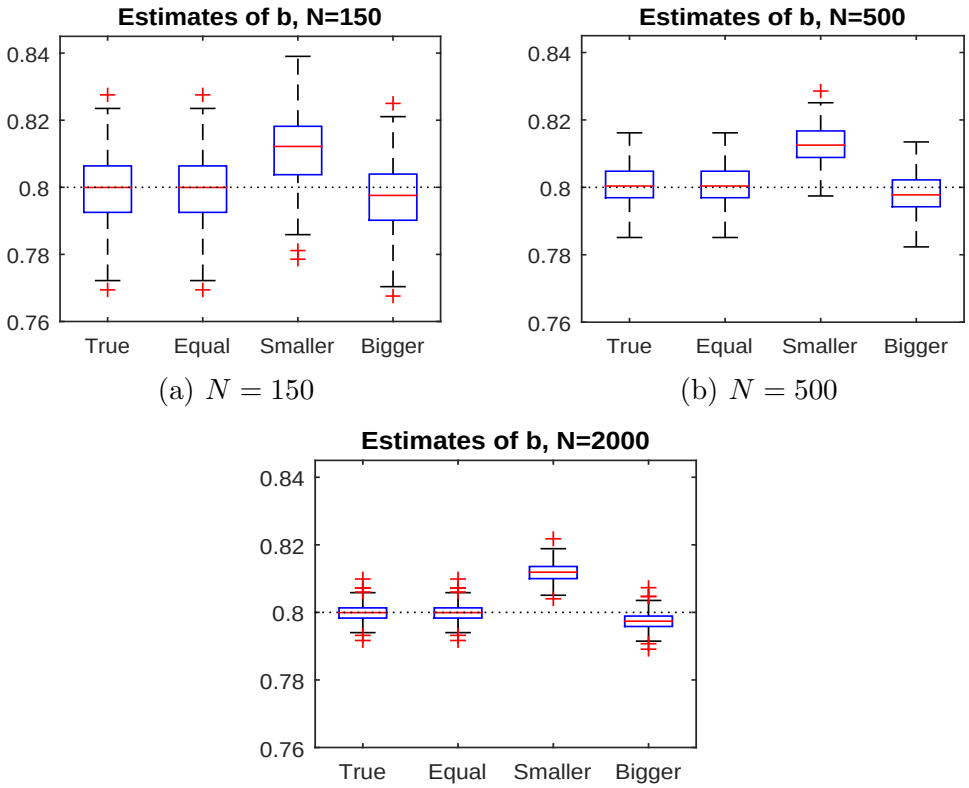
Table 1
Parameter estimation of the slope parameter (b=0.8) under different specifications of weights. The rows "True," "Equal," "Smaller," and "Bigger" show the results for Scenarios 1-4, respectively. The 95% confidence interval is obtained as the sample mean plus/minus twice its standard error calculated from the simulation. The results are based on 200 simulated datasets for each of the four scenarios, and each of the three values of N.

N=150	Mean	Median	MSE	95%CI
True	0.79945	0.79995	$1.137 \cdot 10^{-4}$	(0.79794, 0.80096)
Equal	0.79945	0.79995	$1.137 \cdot 10^{-4}$	(0.79794, 0.80096)
Smaller	0.81101	0.81218	$2.368 \cdot 10^{-4}$	(0.80949, 0.81254)
Bigger	0.79701	0.79758	$1.215 \cdot 10^{-4}$	(0.79550, 0.79851)
N=500	Mean	Median	MSE	95%CI
True	0.80036	0.80039	$4.043 \cdot 10^{-5}$	(0.79946, 0.80126)
Equal	0.80036	0.80039	$4.043 \cdot 10^{-5}$	(0.79946, 0.80126)
Smaller	0.81242	0.81253	$19.38 \cdot 10^{-5}$	(0.81153, 0.81331)
Bigger	0.79781	0.79777	$4.505 \cdot 10^{-5}$	(0.79691, 0.79871)
N=2000	Mean	Median	MSE	95%CI
True	0.80000	0.79993	$7.635 \cdot 10^{-6}$	(0.79961, 0.80039)
Equal	0.80000	0.79993	$7.635 \cdot 10^{-6}$	(0.79961, 0.80039)
Smaller	0.81194	0.81189	$150.3 \cdot 10^{-6}$	(0.81154, 0.81233)
Bigger	0.79748	0.79740	$13.91 \cdot 10^{-6}$	(0.79709, 0.79787)

Table 1 gives the mean squared error (MSE) and the 95% confidence interval (95% CI) for b. First, it is clear that when the weights satisfy the ratio condition (6), the parameter estimate of b is closest to the true value of b (i.e., it has the smallest MSE). If (6) does not hold, the resulting estimate of b is biased: for Scenario 3 (Smaller), the parameter estimate has large positive bias; conversely, negative bias is observed for Scenario 4 (Bigger). When we increase the regression sample size from $N = 150$ to $N = 2000$, the first two scenarios lead to a more accurate estimate of b. Although the MSEs of the third and fourth scenarios decrease with increasing sample size, these two scenarios still yield a parameter estimate of b with significant bias.

The (approximate) 95% CI for b is given in the last column of Table 1. For Scenarios 1 and 2, the 95% CI always contains the true value $b = 0.8$. For Scenarios 3 and 4, the confidence interval excludes the true value of $b = 0.8$ for all three sample sizes. Figure 1 shows the boxplots of $\hat{b}_{lws}$ under the four different scenarios. A biased estimate of b is observed for both Scenarios 3 and 4, but the bias is clearly worse for Scenario 3 (Smaller).

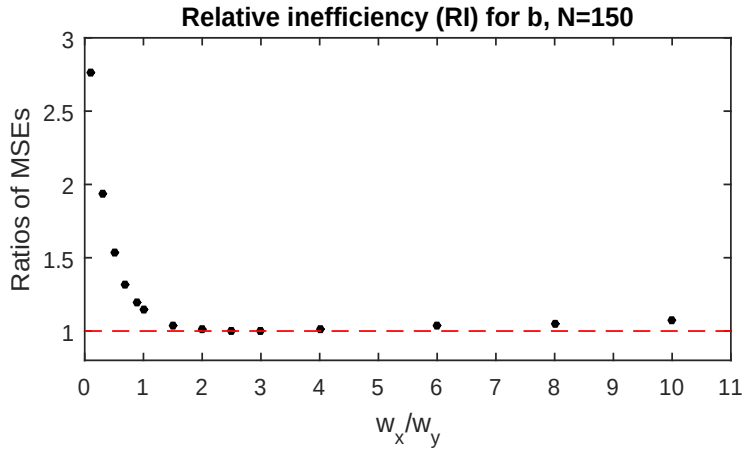
Then we test the sensitivity of $\hat{b}_{lws}$ to the ratio of regression weights using the relative inefficiency (RI), defined as the ratio of the MSE of $\hat{b}_{lws}$ using a ratio of regression weights to that using the actual ratio, $\tilde{\sigma}^2_{y,i}/\tilde{\sigma}^2_{x,i}$. A larger value of RI indicates a worse specification of the ratio of regression weights. The upper panel in Figure 2 shows how the RIs of $\hat{b}_{lws}$ change with different values of the ratio of regression weights, $w_{x,i}/w_{y,i}$. We fixed $w_{x,i}$ at $1/\tilde{\sigma}^2_{x,i}$ and varied the values of $w_{y,i}$ to obtain different ratios of regression weights. The actual ratio in this panel is $\tilde{\sigma}^2_{y,i}/\tilde{\sigma}^2_{x,i} = 3$. It can be seen that the RIs of $\hat{b}_{lws}$ are much more sensitive to Scenario 3 where the ratios have smaller values than the actual ratio of 3, which corroborates our results in Table 1. The lower panel in Figure 2 shows the results for the case $\tilde{\sigma}^2_{y,i}/\tilde{\sigma}^2_{x,i} = 1$, in the simulation described above. Notice that the actual ratio has decreased and, under Scenario 4 where the ratios have bigger values than the actual ratio of 1, the RIs of $\hat{b}_{lws}$ have worsened in the lower panel. Our general conclusion remains however, that underestimation of $\text{var}(Y_i)/\text{var}(X_i)$ will lead to more severely biased estimates of b than its overestimation.



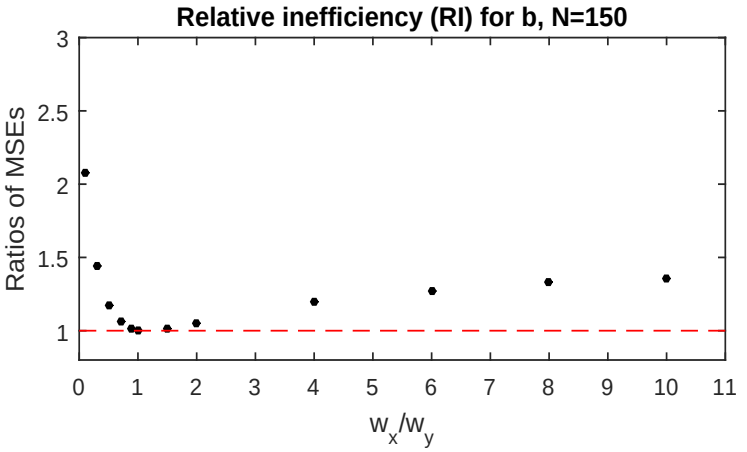
(a) $N = 150$ (b) $N = 500$

(c) $N = 2000$

Fig. 1. Boxplots of the estimates of b under different specifications of weights. The true value of b is 0.8.



(a) $\tilde{\sigma}^2_{y,i} = 1.5, \tilde{\sigma}^2_{x,i} = 0.5$ $(\tilde{\sigma}^2_{y,i}/\tilde{\sigma}^2_{x,i} = 3)$.



(b) $\tilde{\sigma}^2_{y,i} = \tilde{\sigma}^2_{x,i} = 1$ $(\tilde{\sigma}^2_{y,i}/\tilde{\sigma}^2_{x,i} = 1)$.

Fig. 2. The relative inefficiency (defined as the ratio of MSEs) of $\hat{b}_{lws}$ for different values of $w_{x,i}/w_{y,i}$. The dashed line at 1 shows the relative inefficiency of $\hat{b}_{lws}$ using $w_{x,i}/w_{y,i} = \tilde{\sigma}^2_{y,i}/\tilde{\sigma}^2_{x,i}$ $(\tilde{\sigma}^2_{y,i}/\tilde{\sigma}^2_{x,i} = 3$ for the upper panel, and $\tilde{\sigma}^2_{y,i}/\tilde{\sigma}^2_{x,i} = 1$ for the lower panel).

4 Statistical analysis of dependencies in the original TCCON and OCO-2 values

Since both X_i and Y_i are aggregated data based on a dataset of original TCCON observations and a dataset of original OCO-2 observations, respectively, it is necessary to analyze the temporal (for TCCON) and spatial (for OCO-2) dependence structures in order to obtain variances of X_i and Y_i. The TCCON and OCO-2 datasets at Lamont/3590 (which refer to the i-th station/orbit combination) are used to illustrate our methodology.

4.1 TCCON data analysis

There are on the order of 25 TCCON stations in the world and, in what follows, we have chosen the Lamont station located in Oklahoma, USA, to illustrate the appropriate calculation of $\tilde{\sigma}^2_{x,i} = \text{var}(X_i)$ and $\tilde{\sigma}^2_{y,i} = \text{var}(Y_i)$. The Lamont station is in the Southern Great Plains, which has been widely studied in climate-model-calibration contexts [19]. On orbit 3590, the OCO-2 satellite was in target mode, obtaining observations around the Lamont station during a time interval of a few minutes. In the analysis, the mean target time was first obtained as the average of OCO-2's target-start-time and target-end-time; then, as many as 65 TCCON observations, in the time window of approximately ± 1 hour centered at the mean target time, were selected for statistical analysis. The left panel of Figure 3 shows the selected TCCON data for Lamont/3590; the right panel shows the locations of the OCO-2 data from target-start-time to target-end-time.

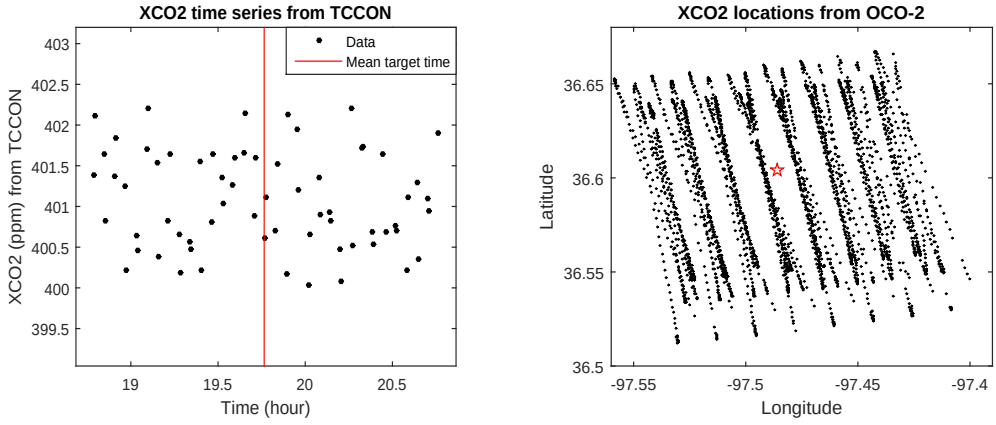


Fig. 3. The TCCON time series and OCO-2 data locations for Lamont/3590. Left panel: TCCON observations versus time, where the vertical line is the mean target time. Right panel: OCO-2 observation locations, where the star shows the location of the Lamont TCCON station.

Since TCCON data at a given ground-monitoring station are observed over time, we model TCCON observations as realizations from a temporal stochastic process (i.e., a time series). For the i-th station/orbit (here, Lamont/3590), let $\{X_{i,1}, \ldots, X_{i,n_{x,i}}\}$ be the $n_{x,i}$ TCCON observations selected in the time window described above, and let $\{t_{i,1}, \ldots, t_{i,n_{x,i}}\}$ be their corresponding observation times. We can expect $n_{x,i} = 65$, although with missing data it may be less. We model $\{X_{i,j}\}_{j=1}^{n_{x,i}}$ as follows:

$$X_{i,j} = x_i + \epsilon_{x,i,j}, \tag{7}$$

where x_i is a fixed but unknown constant (in time) mean parameter, and $\epsilon_{x,i,j}$ is a measurement-error term that we assume to be Gaussian with mean zero and temporal covariance function, $\mathcal{C}_x(\cdot,\cdot;\boldsymbol{\theta}_{x,i})$. In what follows, we capture the temporal dependence through the exponential covariance function,

$$\text{cov}(\epsilon_{x,i,j},\epsilon_{x,i,\ell}) = \mathcal{C}_x(t_{i,j},t_{i,\ell};\boldsymbol{\theta}_{x,i}) = \sigma^2_{x,i}\exp(-|t_{i,j}-t_{i,\ell}|/\phi_{x,i}), \tag{8}$$

where $\boldsymbol{\theta}_{x,i} = \{\sigma^2_{x,i},\phi_{x,i}\}$, and $\sigma^2_{x,i} > 0$ and $\phi_{x,i} > 0$ are the variance and range parameters, respectively. In engineering applications, $\phi_{x,i}$ is sometimes called the e-folding time, and $3\phi_{x,i}$ is sometimes referred to as the equivalent range. For some TCCON stations/orbits, the time series might indicate a non-constant trend over time. We remark that the covariance function can help capture small but apparent departures from a constant trend; if the trend component around the mean target time is not constant, then the range parameter increases to capture the temporal trend in the data.

By using REstricted Maximum Likelihood (REML) estimation for $\sigma^2_{x,i}$ and $\phi_{x,i}$ (see [20,21]), we obtain estimators of covariance parameters that are less biased than those obtained from Maximum Likelihood (ML) estimation. Let $P = \frac{1}{n_{x,i}}\mathbf{1}_{n_{x,i}}\mathbf{1}^T_{n_{x,i}}$ be a projection matrix, where $\mathbf{1}_{n_{x,i}}$ is a column vector of $n_{x,i}$ ones; then the REML approach performs ML estimation on the transformed data, $\tilde{\mathbf{X}}_i = (I_{n_{x,i}} - P)\mathbf{X}_i = (I_{n_{x,i}} - P)\boldsymbol{\epsilon}_{x,i}$, where $\mathbf{X}_i = (X_{i,1},\ldots,X_{i,n_{x,i}})^T$ and $\boldsymbol{\epsilon}_{x,i} = (\epsilon_{x,i,1},\ldots,\epsilon_{x,i,n_{x,i}})^T$.

The left panel in Figure 4 shows the empirical semivariogram (e.g., see [22]) and the fitted semivariogram using REML parameter estimates; the fitted semivariogram values match the empirical ones well. We can see that the fitted semivariogram reaches its sill quickly with increasing time lags, indicating very weak temporal dependence in the TCCON data. Based on the estimates given in Section 5, the equivalent range is $3\hat{\phi}_{x,i} \simeq 1.3\cdot 10^{-3}$ hour, which is much smaller than the time window of 2 hours.

4.2 OCO-2 data analysis

The OCO-2 dataset that targets the TCCON ground station during a given orbit has spatio-temporal locations in a small spatial region ($\sim 0.2\times 0.2$ degrees) within a few minutes (~ 5 minutes). Therefore, a spatial-constant-mean assumption is likely to hold. Recall that the right panel in Figure 3 shows the locations of the OCO-2 data from target-start-time to target-end-time for Lamont/3590; eight footprints are clearly observed (and expected).

Since the time interval of the OCO-2 observations is very short and the corresponding spatial locations are changing during the short time interval, we

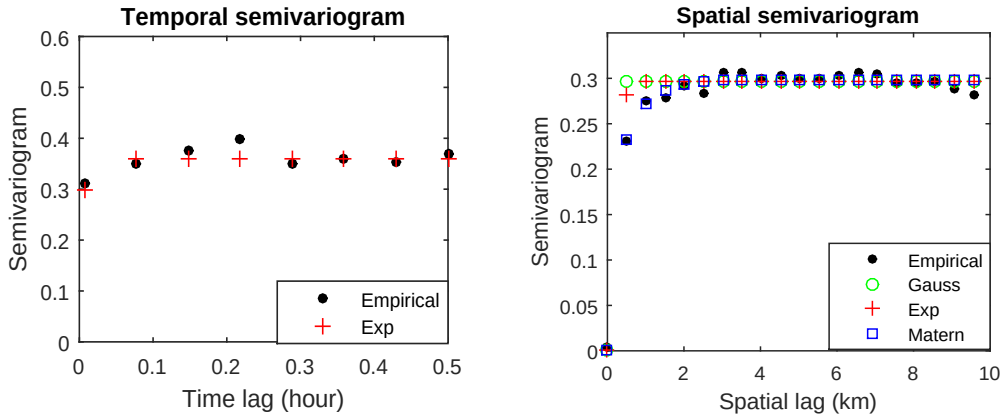


Fig. 4. Empirical and fitted semivariogram plots for Lamont/3590. Left panel: semivariograms for the TCCON (temporal semivariograms) dataset. Right panel: semivariograms for the OCO-2 (spatial isotropic semivariograms) dataset.

model the OCO-2 observations using a purely spatial process. Let $\{Y_{i,1}, \ldots, Y_{i,n_{y,i}}\}$ be the $n_{y,i}$ OCO-2 observations for the i-th station/orbit, with corresponding spatial locations $\{\mathbf{s}_{i,1}, \ldots, \mathbf{s}_{i,n_{y,i}}\}$. At the resolution of 15km, the chordal distance, which is a Euclidean distance in three-dimensional space, is used to quantify "closeness" of the OCO-2 data locations.

We model $\{Y_{i,j} : j = 1, \ldots, n_{y,i}\}$ as realizations from a spatial Gaussian process, as follows:

$$Y_{i,j} = y_i + \epsilon_{y,i,j}, \tag{9}$$

where y_i is a fixed but unknown spatially constant mean parameter, and $\epsilon_{y,i,j}$ is a measurement-error term that we assume to be Gaussian with mean zero and spatial covariance function, $\mathcal{C}_y(\cdot, \cdot; \boldsymbol{\theta}_{y,i})$. Let $\mathcal{C}_y(\cdot, \cdot; \boldsymbol{\theta}_{y,i})$ be the flexible isotropic Matérn covariance function (see [22,23]):

$$\mathcal{C}_y(\mathbf{s}_{i,j}, \mathbf{s}_{i,\ell}; \boldsymbol{\theta}_{y,i}) = \frac{\sigma^2_{y,i} 2^{1-\nu_{y,i}}}{\Gamma(\nu_{y,i})} \left(\frac{\|\mathbf{s}_{i,j} - \mathbf{s}_{i,\ell}\|}{\phi_{y,i}} \right)^{\nu_{y,i}} \mathcal{K}_{\nu_{y,i}} \left(\frac{\|\mathbf{s}_{i,j} - \mathbf{s}_{i,\ell}\|}{\phi_{y,i}} \right), \tag{10}$$

where $\boldsymbol{\theta}_{y,i} = \{\sigma^2_{y,i}, \phi_{y,i}, \nu_{y,i}\}$, $\sigma^2_{y,i} > 0$ is the variance parameter, $\phi_{y,i} > 0$ is the range parameter, and $\nu_{y,i} > 0$ is the smoothness parameter. In (10), $\Gamma(\cdot)$ is the gamma function, and $\mathcal{K}_{\nu_{y,i}}(\cdot)$ is a modified Bessel function of the second kind. The covariance model given by (10) provides extra flexibility for modeling the smoothness of the process with the inclusion of the third parameter, $\nu_{y,i}$. The exponential covariance function in (8) is a special case of the Matérn model with $\nu_{y,i} = 0.5$; when $\nu_{y,i}$ tends to infinity, the so-called Gaussian covariance function is obtained.

Outliers are typical in most remote sensing datasets; thus, the estimates of covariance-model parameters need to be robust to the presence of outliers. We therefore use robust estimators of the semivariogram (e.g., [22,24]) and fit covariance-function parameters using weighted least squares [22].

The Cressie-Hawkins semivariogram estimator [22] is:

$$\hat{\gamma}(\mathbf{h}(k)) = \frac{1}{2}\left(\frac{1}{|N(\mathbf{h}(k))|}\sum_{N(\mathbf{h}(k))}|Y_{i,j}-Y_{i,\ell}|^{1/2}\right)^4 \Big/ \left(0.457+\frac{0.494}{|N(\mathbf{h}(k))|}\right), \quad (11)$$

where $N(\mathbf{h}(k)) \equiv \{(j,\ell) : \mathbf{s}_{i,j}-\mathbf{s}_{i,\ell} \in \text{tol}(\mathbf{h}(k)), \text{and } j,\ell = 1,\ldots,n_{y,i}\}$, $\text{tol}(\mathbf{h}(k))$ is a pre-specified tolerance region around the spatial lag $\mathbf{h}(k)$, and $|N(\mathbf{h}(k))|$ is the number of distinct pairs in $N(\mathbf{h}(k))$. Then the parameters $\boldsymbol{\theta}_{y,i}$ are estimated by weighted least squares (see [22]); that is, we minimize with respect to $\boldsymbol{\theta}_{y,i}$,

$$W(\boldsymbol{\theta}_{y,i}) = \sum_{k=1}^{K}|N(\mathbf{h}(k))|\left(\frac{\hat{\gamma}(\mathbf{h}(k))}{\gamma(\mathbf{h}(k);\boldsymbol{\theta}_{y,i})}-1\right)^2. \quad (12)$$

In (12), $\gamma(\mathbf{h}(k);\boldsymbol{\theta}_{y,i}) = \mathcal{C}_y(\mathbf{0};\boldsymbol{\theta}_{y,i}) - \mathcal{C}_y(\mathbf{h}(k);\boldsymbol{\theta}_{y,i})$ is the semivariogram based on the covariance model in (10). Notice that $\gamma(\cdot;\boldsymbol{\theta}_{y,i})$ does not depend on the spatial mean y_i, and hence neither does (12). In (12), we choose K initial lags, where for $k = 1,\ldots,K$, $\|\mathbf{h}(k)\| \leq \frac{1}{2}\max\{\|\mathbf{s}_{i,j}-\mathbf{s}_{i,k}\| : j,k = 1,\ldots,n_{y,i}\}$, and we adjust $\text{tol}(\mathbf{h}(k))$ such that $|N(\mathbf{h}(k))|$ does not fall below 30; these are empirical rules of thumb that work well (e.g., see [25,26]). In our application to the Lamont/3590 OCO-2 dataset, $K = 20$ equally spaced spatial lags were used to evaluate $W(\boldsymbol{\theta}_{y,i})$.

We fitted three Matérn models with different smoothness parameters, namely a Matérn model with $\hat{\nu}_{y,i}$ fitted via minimizing (12); the exponential model ($\nu_{y,i} = 1/2$), $\sigma^2_{y,i}\exp(-\|\mathbf{s}_{i,j}-\mathbf{s}_{i,k}\|/\phi_{y,i})$; and the Gaussian model ($\nu_{y,i}\to\infty$), $\sigma^2_{y,i}\exp(-\|\mathbf{s}_{i,j}-\mathbf{s}_{i,k}\|^2/\phi_{y,i})$. The right panel of Figure 4 shows the empirical semivariograms versus the fitted semivariograms obtained from these three models. It is clear that the fitted Matérn covariance model is the best, due to its ability to capture the smoothness of the spatial process. The value of $W(\boldsymbol{\theta}_{y,i})$ using the fitted Matérn model is 687.989, which is much smaller than the fitted exponential model's value of 1261.964 and than the fitted Gaussian model's value of 1384.135.

The empirical semivariogram of the OCO-2 data attains its sill gently as a function of spatial lag, indicating quite strong spatial dependence. Based on the parameter estimation results in Section 5, the equivalent range is about 1.37km. Compared with the spatial domain of approximately 15km $\times$ 15km,

the correlations among the OCO-2 observations are substantial, and hence they will have an effect on $\tilde{\sigma}^2_{y,i} = \text{var}(Y_i)$ and on $\tilde{n}_{y,i}$, the effective sample size.

5 Calculation of X_i and Y_i and of their associated variances

In this section, we discuss how to estimate the variances of the measurement errors of X_i in (7) and Y_i in (9), which are then used to define the weights in (3). Use of inappropriate values of $\tilde{\sigma}^2_{x,i}$ and $\tilde{\sigma}^2_{y,i}$ may lead to biased estimates of b. The TCCON and OCO-2 datasets at Lamont/3590 (here index i) [27] are used to illustrate the appropriate calculation of the all-important regression weights, $w_{x,i} = 1/\tilde{\sigma}^2_{x,i}$ and $w_{y,i} = 1/\tilde{\sigma}^2_{y,i}$.

5.1 Variance estimation in the TCCON dataset

Generally speaking, TCCON datasets are of very high quality, with very few outliers. Consequently, the sample mean, $X_i \equiv \frac{1}{n_{x,i}} \sum_{j=1}^{n_{x,i}} X_{i,j}$, serves as the i-th representative point for TCCON in the regression analysis. From the model (7), $X_i = x_i + \bar{\epsilon}_{x,i}$, where $\bar{\epsilon}_{x,i} = \frac{1}{n_{x,i}} \sum_{j=1}^{n_{x,i}} \epsilon_{x,i,j}$; hence,

$$\text{var}(X_i) = \text{var}(\bar{\epsilon}_{x,i}) = \frac{\sigma^2_{x,i}}{n_{x,i}} + \frac{1}{n^2_{x,i}} \sum_{j=1}^{n_{x,i}} \sum_{k \neq j} C_{x,i;j,k}, \tag{13}$$

where $C_{x,i;j,k} = \mathcal{C}_x(t_{i,j}, t_{i,k}; \boldsymbol{\theta}_{x,i})$. By substituting into (13) the REML estimates $\hat{\boldsymbol{\theta}}_{x,i} = \{\hat{\sigma}^2_{x,i}, \hat{\phi}_{x,i}\}$ (Section 4.1) of temporal covariance function parameters $\boldsymbol{\theta}_{x,i}$, $\text{var}(X_i)$ can be readily obtained.

If the measurement errors $\{\epsilon_{x,i,j}\}_{j=1}^{n_{x,i}}$ are independent, then $\text{var}(X_i) = \sigma^2_{x,i}/n_{x,i}$. In the presence of temporal dependence, the effective sample size, $\tilde{n}_{x,i}$ (in contrast to the actual sample size), for computing the variance of X_i is defined as (see [22], pp. 14-15):

$$\tilde{n}_{x,i} = \sigma^2_{x,i}/\text{var}(\bar{\epsilon}_{x,i}) = \left(\frac{1}{n^2_{x,i}} \sum_{j=1}^{n_{x,i}} \sum_{k=1}^{n_{x,i}} \rho_{x,i;j,k} \right)^{-1}, \tag{14}$$

where $\rho_{x,i;j,k} = C_{x,i;j,k}/\sigma^2_{x,i}$. Formula (14) depends on $\sigma^2_{x,i}$ and $\phi_{x,i}$; estimates $\hat{\sigma}^2_{x,i}$ and $\hat{\phi}_{x,i}$ are substituted into (14) to obtain the final result. The effective sample size (14) is always smaller than the actual sample size when the measurement errors are positively correlated. For Lamont/3590, Table 2 shows

that $\tilde{n}_{x,i}$ is very close to $n_{x,i}$, since there is only weakly positive temporal dependence in the TCCON dataset.

5.2 Variance estimation in the OCO-2 dataset

OCO-2 measurements are based on reflected energy from Earth's surface that has traveled through the atmosphere twice. This, and the impact of environmental factors such as clouds and aerosols on the data-retrieval process, result in high variability and a number of outliers in the OCO-2 dataset. Hence, rather than the sample mean, the sample median given by $Y_i = \text{median}\{Y_{i,j} : j = 1, \ldots, n_{y,i}\} \equiv \text{med}(Y_{i,j})$ is chosen as a robust OCO-2 representative point in the pair (X_i, Y_i) used in the regression. The OCO-2 data are spatially correlated, so we need to specify the (approximate) variance of the sample median under this dependence; detailed calculations are given in the Appendix.

From the model (9), $Y_i = y_i + \text{med}(\epsilon_{y,i,j})$, and hence $\text{var}(Y_i) = \text{var}(\text{med}(\epsilon_{y,i,j}))$. Under mild conditions that are given in [28,29], the large-sample variance is,

$$\text{var}(Y_i) = \text{var}(\text{med}(\epsilon_{y,i,j})) \simeq \frac{\pi\sigma^2_{y,i}}{2n_{y,i}} + \frac{\sigma^2_{y,i}}{n^2_{y,i}} \sum_{j=1}^{n_{y,i}} \sum_{k\neq j} \arcsin(C_{y,i;j,k}/\sigma^2_{y,i}), \quad (15)$$

where $C_{y,i;j,k} = \text{cov}(\epsilon_{y,i,j}, \epsilon_{y,i,k}) = \mathcal{C}_y(\mathbf{s}_{i,j}, \mathbf{s}_{i,k}; \boldsymbol{\theta}_{y,i})$. By substituting into (15) the semivariogram-based weighted-least-squares estimates, $\hat{\boldsymbol{\theta}}_{y,i} = \{\hat{\sigma}^2_{y,i}, \hat{\phi}_{y,i}, \hat{\nu}_{y,i}\}$ (Section 4.2) of spatial covariance parameters $\boldsymbol{\theta}_{y,i}$, $\text{var}(Y_i)$ can be readily obtained.

If the measurement errors $\{\epsilon_{y,i,j}\}_{j=1}^{n_{y,i}}$ are independent, then $\text{var}(Y_i) \simeq \frac{\pi}{2}(\sigma^2_{y,i}/n_{y,i})$. Consequently, the effective sample size, $\tilde{n}_{y,i}$, for computing the variance of Y_i is,

$$\tilde{n}_{y,i} = \frac{\pi}{2} \cdot \frac{\sigma^2_{y,i}}{\text{var}(\text{med}(\epsilon_{y,i,j}))} = \frac{\pi}{2} \left(\frac{1}{n^2_{y,i}} \sum_{j=1}^{n_{y,i}} \sum_{k=1}^{n_{y,i}} \arcsin(C_{y,i,j,k}/\sigma^2_{y,i}) \right)^{-1}, \quad (16)$$

which accounts for the reduction of the actual sample size for calculating the variance of $\text{var}(Y_i)$ under dependence. Formula (16) depends on $\sigma^2_{y,i}$, $\phi_{y,i}$, and $\nu_{y,i}$; estimates $\hat{\sigma}^2_{y,i}$, $\hat{\phi}_{y,i}$, and $\hat{\nu}_{y,i}$ are substituted into (16) to obtain the final result.

For Lamont/3590, Table 2 shows that the effective sample size $\tilde{n}_{y,i}$ is much smaller than the actual sample size $n_{y,i}$, since there is strongly positive spatial dependence in the OCO-2 dataset. Similar calculations can be made for flux

inversions, when variances are needed for spatially aggregated mole-fraction data.

5.3 Results for Lamont/3590

Table 2 shows the results of parameter estimation and the calculation of effective sample sizes for both the TCCON and the OCO-2 datasets associated with Lamont/3590. For the TCCON dataset, the range-parameter estimate $\hat{\phi}_{x,i}$ is very small relative to the 2-hour time window over which the data were collected, and so the temporal correlations drop quickly with increasing time lags. The weak correlations in the TCCON data lead to an effective sample size, $\tilde{n}_{x,i}$, that is very close to the actual sample size $n_{x,i}$. For the OCO-2 dataset, the range-parameter estimate $\hat{\phi}_{y,i}$ is much larger, relative to the 15km $\times$ 15km spatial window in which the data were collected. Hence, the strongly positive spatial correlations in the OCO-2 data result in an effective sample size, $\tilde{n}_{y,i}$, that is much smaller than the actual sample size $n_{y,i}$.

Table 2
Parameter estimation results and the effective sample sizes for Lamont/3590.

TCCON	X_i	$\tilde{\sigma}^2_{x,i}$	$S^2_{x,i}$	$n_{x,i}$	$\tilde{n}_{x,i}$	$\hat{\sigma}^2_{x,i}$	$\hat{\phi}_{x,i}$	$\hat{\nu}_{x,i}$
	401.0840	0.0063	0.3602	65	57.05	0.3607	0.0039	0.5(fixed)
OCO-2	Y_i	$\tilde{\sigma}^2_{y,i}$	$S^2_{y,i}$	$n_{y,i}$	$\tilde{n}_{y,i}$	$\hat{\sigma}^2_{y,i}$	$\hat{\phi}_{y,i}$	$\hat{\nu}_{y,i}$
	400.0395	0.0023	0.3025	2961	202.32	0.2989	0.7117	0.1849

Based on our calculations from (6), we obtain the relative regression weights of $w_{x,i}/w_{y,i} = \tilde{\sigma}^2_{y,i}/\tilde{\sigma}^2_{x,i} = 0.3651$. For version 7 of the OCO-2 retrieval data product, the calculation used $w_{x,i}/w_{y,i} = S^2_{y,i}/S^2_{x,i} = 0.8398$, where $S^2_{x,i}$ and $S^2_{y,i}$ are sample variances of the individual TCCON and OCO-2 datasets, respectively. The sample variances were used to define weights for version 7 of the OCO-2 retrieval data product, because they better reflect variability observed in the data than the squared standard errors. Based on our calculations in this article, the sample-variance-based weights used for version 7 overestimate the approximate ratio of weights for Lamont/3590 (i.e., Scenario 4, which is "Bigger").

6 Concluding remarks

In this paper, we have proposed a statistical procedure for obtaining regression weights that lead to consistent estimation of regression coefficients. The

application is to calibration of satellite remote sensing observations obtained by looking down from space, calibrated to ground-based observations obtained by looking up from the ground. Specifically, OCO-2 values $\{Y_i\}$ are regressed on TCCON values $\{X_i\}$. In this article, we show that the appropriate regression weights depend on temporal (TCCON) and spatial (OCO-2) dependence structures. Specification of the regression weights associated with X_i and Y_i are crucial for obtaining an (asymptotically) unbiased, least-sum-of-weighted-squares estimator, $\hat{b}_{lws}$. When $w_{x,i}/w_{y,i} = \tilde{\sigma}^2_{y,i}/\tilde{\sigma}^2_{x,i}$, the estimating equation for b is unbiased, which results in (asymptotically) unbiasedness of $\hat{b}_{lws}$. Therefore, it is desirable to use unbiased estimates of variances of X_i and Y_i for defining the weights. Since X_i and Y_i are aggregated data calculated based on sets of individual TCCON and OCO-2 observations, respectively, we explore the temporal-dependence and spatial-dependence structures in the TCCON and OCO-2 datasets for estimating $\tilde{\sigma}^2_{x,i}$ and $\tilde{\sigma}^2_{y,i}$, respectively. Based on our analysis, the individual observations in the TCCON dataset are weakly correlated in time, resulting in an effective sample size very close to the actual sample size; in contrast, the individual observations in the OCO-2 dataset have nonnegligible correlations in space, resulting in an effective sample size much smaller than the actual sample size. Our results show that any bias correction to produce a new version of the OCO_2 data should use regression weights that are statistically determined.

Version 7 used TCCON and OCO-2 datasets at $N = 66$ station/orbit combinations. Future work will result in a careful analysis of all these datasets, which can then be used to obtain consistent estimators, $\hat{a}_{lws}$ and $\hat{b}_{lws}$. This work will build on the substantial methodology given in the preceding sections and the appendix.

Acknowledgements
The OCO-2 data used in this article were produced by the OCO-2 project at the Jet Propulsion Laboratory, California Institute of Technology, and obtained from the OCO-2 data archive maintained at the NASA Goddard Earth Science Data and Information Services Center (`http://disc.sci.gsfc.nasa.gov/OCO-2`). TCCON data were obtained from the TCCON Data Archive, hosted by the Carbon Dioxide Information Analysis Center (CDIAC) (`tccon.ornl.gov`). The Lamont TCCON station is funded by NASA grants NNX14AI60G, NNX11AG01G, NAG5-12247, NNG05-GD07G, and NASA's Orbiting Carbon Observatory Program. We are grateful to the DOE ARM program for technical support at Lamont. Zhang and Cressie's research was partially supported by a 2015-2017 Australian Research Council Discovery Grant, number DP150104576; Cressie's research was also partially supported by NASA grant NNH11-ZDA001N-OCO2. We would like to thank Paul Wennberg and Camille Viatte for early discussions and their interest in this research.

References

[1] D. Crisp, et al., The on-orbit performance of the Orbiting Carbon Observatory (OCO-2), in prep. (2016).

[2] R. K. Pachauri, M. Allen, V. Barros, J. Broome, W. Cramer, R. Christ, J. Church, L. Clarke, Q. Dahe, P. Dasgupta, et al., Climate Change 2014: Synthesis Report. Contribution of Working Groups i, ii and iii to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Tech. rep. (2014).

[3] C. L. Quéré, R. J. Andres, T. Boden, T. Conway, R. Houghton, J. I. House, G. Marland, G. P. Peters, G. Van der Werf, A. Ahlström, et al., The global carbon budget 1959–2011, Earth System Science Data 5 (2013) 165–185.

[4] C. Le Quéré, M. R. Raupach, J. G. Canadell, G. Marland, L. Bopp, P. Ciais, T. J. Conway, S. C. Doney, R. A. Feely, P. Foster, et al., Trends in the sources and sinks of carbon dioxide, Nature Geoscience 2 (2009) 831–836.

[5] S. Boland, H. Bösch, L. Brown, J. Burrows, P. Ciais, B. Connor, D. Crisp, S. Denning, S. Doney, R. Engelen, et al., The need for atmospheric carbon dioxide measurements from space: Contributions from a rapid reflight of the Orbiting Carbon Observatory, Tech. rep., OCO Science Team, NASA Jet Propulsion Labaratory (2009).

[6] C. Miller, D. Crisp, P. DeCola, S. Olsen, J. T. Randerson, A. M. Michalak, A. Alkhaled, P. Rayner, D. J. Jacob, P. Suntharalingam, et al., Precision requirements for space-based data, Journal of Geophysical Research: Atmospheres 112 (D10) (2007) n/a–n/a. doi:10.1029/2006JD007659.

[7] D. Wunch, G. C. Toon, J.-F. L. Blavier, R. A. Washenfelder, J. Notholt, B. J. Connor, D. W. Griffith, V. Sherlock, P. O. Wennberg, The total carbon column observing network, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 369 (2011) 2087–2112.

[8] D. Wunch, G. C. Toon, P. O. Wennberg, S. C. Wofsy, B. B. Stephens, M. L. Fischer, O. Uchino, J. B. Abshire, P. Bernath, S. C. Biraud, et al., Calibration of the Total Carbon Column Observing Network using aircraft profile data, Atmospheric Measurement Techniques 3 (2010) 1351–1362.

[9] D. Wunch, G. Osterman, B. Fisher, B. Naylor, C. M. Roehl, C. O'Dell, et al., Comparisons of the Orbiting Carbon Observatory-2 (OCO-2) XCO2 measurements with TCCON, in prep. (2016).

[10] L. Mandrake, C. O'Dell, D. Wunch, P. O. Wennberg, B. Fisher, G. B. Osterman, A. Eldering, Orbiting Carbon Observatory–2 (OCO-2) warn level, bias correction, and lite file product description, Tech. rep., JPL (2015).

[11] C. W. O'Dell, P. O. Wennberg, A. Eldering, D. Crisp, M. R. Gunson, B. Fisher, The OCO-2 retrieval algorithm, in prep. (2016).

[12] J. Wennberg, A. Gittelsohn, Small area variations in health care delivery, Science 182 (1973) 1102–1108.

[13] Goddard Earth Sciences Data and Information Services Center OCO-2 Data Holdings (2016). doi:http://disc.sci.gsfc.nasa.gov/OCO-2.

[14] D. York, Least-squares fitting of a straight line, Canadian Journal of Physics 44 (1966) 1079–1086.

[15] D. York, N. M. Evensen, M. L. Martınez, J. D. B. Delgado, Unified equations for the slope, intercept, and standard errors of the best straight line, American Journal of Physics 72 (2004) 367–375.

[16] D. York, Least squares fitting of a straight line with correlated errors, Earth and Planetary Science Letters (1968) 320–324.

[17] V. P. Godambe, An optimum property of regular maximum likelihood estimation, Annals of Mathematical Statistics 31 (1960) 1208–1211.

[18] G. Y. Yi, N. Reid, A note on mis-specified estimating functions, Statistica Sinica 20 (2010) 1749–1769.

[19] B. Yang, Y. Qian, G. Lin, R. Leung, Y. Zhang, Some issues in uncertainty quantification and parameter tuning: A case study of convective parameterization scheme in the WRF regional climate model, Atmospheric Chemistry and Physics 12 (2012) 2409–2427.

[20] H. D. Patterson, R. Thompson, Recovery of inter-block information when block sizes are unequal, Biometrika 58 (1971) 545–554.

[21] D. A. Harville, Maximum likelihood approaches to variance component estimation and to related problems, Journal of the American Statistical Association 72 (1977) 320–338.

[22] N. Cressie, Statistics for Spatial Data, revised edition, John Wiley & Sons, Hoboken, NJ, 1993.

[23] B. Matérn, Spatial Variation, Vol. 36, Springer Science & Business Media, New York, NY, 2013.

[24] N. Cressie, D. M. Hawkins, Robust estimation of the variogram: I, Journal of the International Association for Mathematical Geology 12 (1980) 115–125.

[25] A. G. Journel, C. J. Huijbregts, Mining Geostatistics, Academic Press, London, UK, 1978.

[26] N. Cressie, A graphical procedure for determining nonstationarity in time series, Journal of the American Statistical Association 83 (1988) 1108–1116.

[27] P. O. Wennberg, D. Wunch, C. Roehl, J.-F. Blavier, G. C. Toon, N. Allen, P. Dowell, K. Teske, C. Martin, J. Martin, TCCON data from Lamont, Oklahoma, USA (2014). doi:10.14291/tccon.ggg2014.lamont01.R0/1149159.

[28] N. Cressie, G. Glonek, Median based covariogram estimators reduce bias, Statistics & Probability Letters 2 (1984) 299–304.

[29] P. K. Sen, On the Bahadur representation of sample quantiles for sequences of φ-mixing random variables, Journal of Multivariate Analysis 2 (1972) 77–95.

[30] I. Abrahamson, Orthant probabilities for the quadrivariate normal distribution, Annals of Mathematical Statistics 35 (1964) 1685–1703.

APPENDIX . Approximate variance of the sample median under dependence

We provide some details for obtaining the approximate variance of the sample median under dependence. Since many of the target-mode OCO-2 datasets contain outliers, the sample median replaces the sample mean as the representative value, Y_i, for the i-th station/orbit. Recall that we model an OCO-2 dataset as a realization from a Gaussian process with a constant mean y_i and an isotropic spatial covariance function, $\mathcal{C}_y(\cdot,\cdot;\boldsymbol{\theta}_{y,i})$ given by (10). Under mild conditions, the sample median converges almost surely to y_i and, to leading order, we may write the sample median as (see [28,29]),

$$\tilde{Y}_i = y_i + \frac{1}{n_{y,i}} \sum_{j=1}^{n_{y,i}} \mathrm{sgn}(Y_{i,j} - y_i)/(2f(y_i)).$$

In the equation above, the function, $\mathrm{sgn}(x)$, is a sign function such that $\mathrm{sgn}(x) = 1$ if $x > 0$, $\mathrm{sgn}(x) = 0$ if $x = 0$, and $\mathrm{sgn}(x) = -1$ if $x < 0$; and $f(\cdot)$ is the density function of $Y_{i,j}$, which here is Gaussian.

Therefore, the asymptotic variance of the sample median is, to leading order,

$$\begin{aligned}
\mathrm{var}(\tilde{Y}_i) &= \frac{1}{n_{y,i}^2} \sum_{j=1}^{n_{y,i}} \sum_{k=1}^{n_{y,i}} \mathrm{cov}(\mathrm{sgn}(Y_{i,j} - y_i), \mathrm{sgn}(Y_{i,k} - y_i))/(2f(y_i))^2 \\
&= \frac{1}{n_{y,i}^2} \sum_{j=1}^{n_{y,i}} \sum_{k=1}^{n_{y,i}} E(\mathrm{sgn}(Y_{i,j} - y_i)\mathrm{sgn}(Y_{i,k} - y_i))/(2f(y_i))^2 \\
&= \frac{1}{n_{y,i}^2} \sum_{j=1}^{n_{y,i}} \frac{1}{(2f(y_i))^2} + \frac{1}{n_{y,i}^2} \sum_{j=1}^{n_{y,i}} \sum_{k \neq j} E(\mathrm{sgn}(Y_{i,j} - y_i)\mathrm{sgn}(Y_{i,k} - y_i))/(2f(y_i))^2,
\end{aligned}$$

since $E(\mathrm{sgn}(Y_{i,j})) = 0$ and $E(\mathrm{sgn}(Y_{i,j})^2) = 1$. Now,

$$
\begin{aligned}
E(\text{sgn}(Y_{i,j}-y_i)\text{sgn}(Y_{i,k}-y_i)) = & E(\mathbf{1}_{Y_{i,j}>y_i}\mathbf{1}_{Y_{i,k}>y_i}) - E(\mathbf{1}_{Y_{i,j}>y_i}\mathbf{1}_{Y_{i,k}<y_i}) \\
& -E(\mathbf{1}_{Y_{i,j}<y_i}\mathbf{1}_{Y_{i,k}>y_i}) + E(\mathbf{1}_{Y_{i,j}<y_i}\mathbf{1}_{Y_{i,k}<y_i}) \\
= & P(Y_{i,j}>y_i, Y_{i,k}>y_i) - P(Y_{i,j}>y_i, Y_{i,k}<y_i) \\
& -P(Y_{i,j}<y_i, Y_{i,k}>y_i) + P(Y_{i,j}<y_i, Y_{i,k}<y_i).
\end{aligned}
$$

By Sheppard's theorem in [30], $P(Y_{i,j}>y_i, Y_{i,k}>y_i) = \frac{1}{4}+\frac{1}{2\pi}\text{arcsin}(C_{y,i;j,k}/(C_{y,i;j,j}C_{y,i;k,k})^{1/2})$, where $C_{y,i;j,k}$ is the covariance of $Y_{i,j}$ and $Y_{i,k}$, and $C_{y,i;j,j} = C_{y,i;k,k} = \sigma^2_{y,i}$. Since $P(Y_{i,j}<y_i, Y_{i,k}<y_i) = P(Y_{i,j}>y_i, Y_{i,k}>y_i)$, and

$$
P(Y_{i,j}>y_i, Y_{i,k}<y_i) + P(Y_{i,j}<y_i, Y_{i,k}>y_i) = 1 - 2P(Y_{i,j}>y_i, Y_{i,k}>y_i),
$$

we readily obtain,

$$
E(\text{sgn}(Y_{i,j}-y_i)\text{sgn}(Y_{i,k}-y_i)) = \frac{2}{\pi}\text{arcsin}(C_{y,i;j,k}/(C_{y,i;j,j}C_{y,i;k,k})^{1/2}).
$$

Recall that in the Gaussian case, the marginal density results in $f(y_i) = (2\pi\sigma^2_{y,i})^{-1/2}$, and hence $1/(2f(y_i))^2 = \pi\sigma^2_{y,i}/2$. Therefore, the asymptotic variance of the sample median is,

$$
\begin{aligned}
\text{var}(\text{med}(Y_{i,j})) = & \frac{1}{n_{y,i}}\cdot\frac{\pi\sigma^2_{y,i}}{2} + \frac{1}{n^2_{y,i}}\sum_{j=1}^{n_{y,i}}\sum_{k\neq j}\frac{2}{\pi}\text{arcsin}(C_{y,i;j,k}/\sigma^2_{y,i})\cdot\frac{\pi\sigma^2_{y,i}}{2} \\
= & \frac{\sigma^2_{y,i}}{n_{y,i}}\left(\frac{\pi}{2}\right) + \frac{\sigma^2_{y,i}}{n^2_{y,i}}\sum_{j=1}^{n_{y,i}}\sum_{k\neq j}\text{arcsin}(C_{y,i;j,k}/\sigma^2_{y,i}) \\
= & \frac{\sigma^2_{y,i}}{n^2_{y,i}}\sum_{j=1}^{n_{y,i}}\sum_{k=1}^{n_{y,i}}\text{arcsin}(C_{y,i;j,k}/\sigma^2_{y,i}).
\end{aligned}
$$